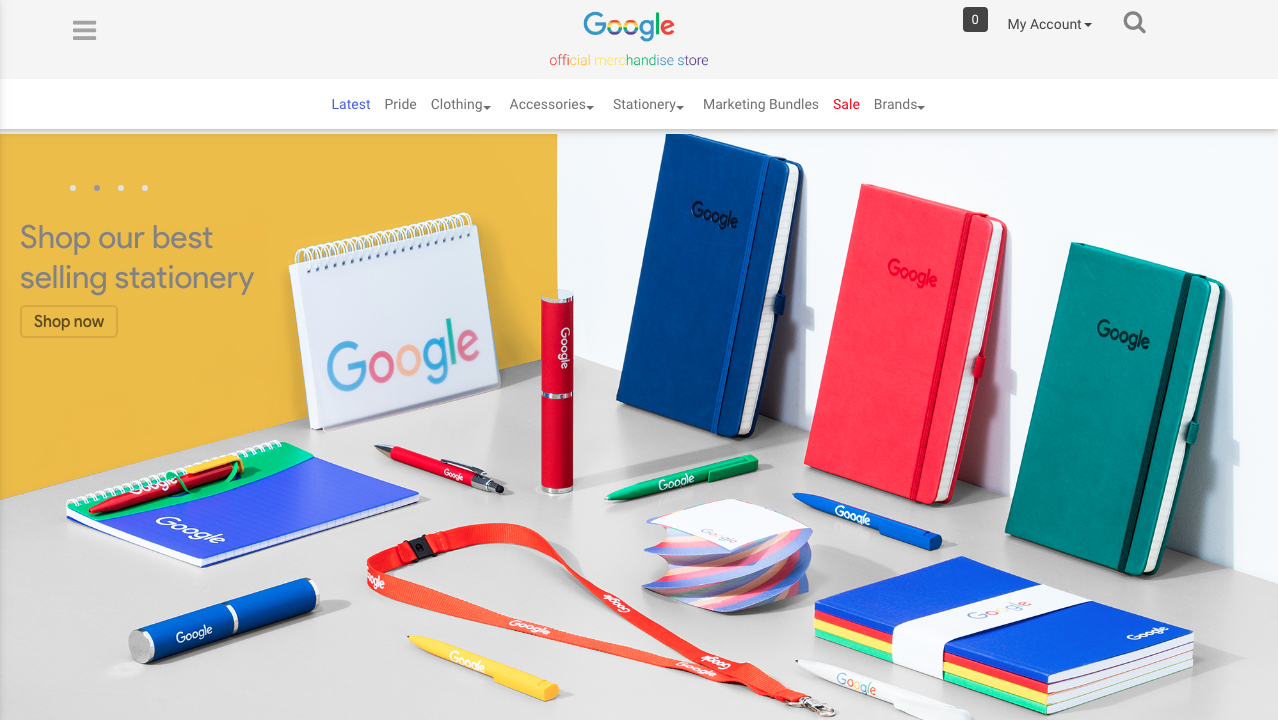
# Analytics & Predictive Modeling

2019 June





## Expected learning outcomes

**Knowledge:**

* Analyse a case of customer revenue for Google Merchandise Store, using Google Analytics data.
* Understand the various features available from Google Analytics data.

**Abilities:**

* Able to conduct exploratory data analysis (EDA) using ***R***
* Able to use predictive models to forecast future customer revenue using ***R***
* Able to identify important features which can indicate customer purchases (data insights)
* Able to build interactive data visualization dashboard to present results & insights using ***Tableau***

## Case Study: Google Analytics Customer Revenue Prediciton

<https://www.kaggle.com/c/ga-customer-revenue-prediction>

The 80/20 rule has proven true for many businesses–only a small percentage of customers produce most of the revenue. As such, marketing teams are challenged to make appropriate investments in promotional strategies.

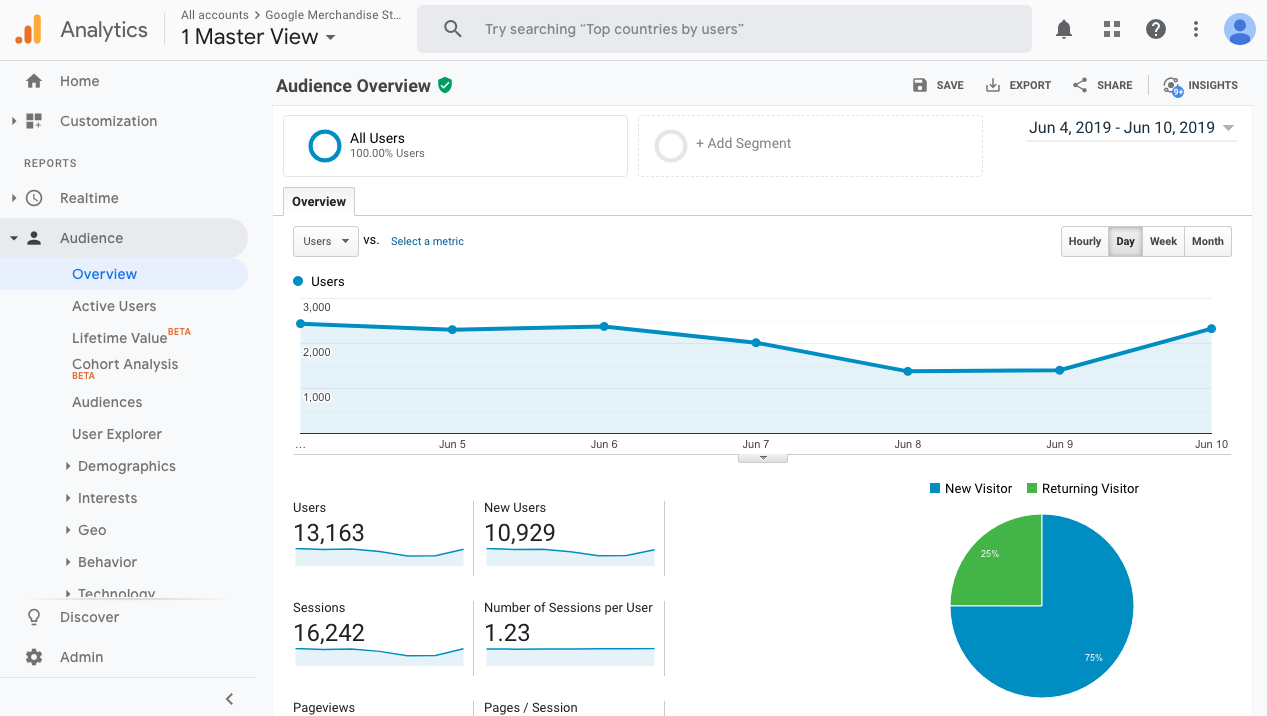
RStudio, the developer of free and open tools for R and enterprise-ready products for teams to scale and share work, has partnered with Google Cloud and Kaggle to demonstrate the business impact that thorough data analysis can have.

In this competition, you’re challenged to analyze a Google Merchandise Store (also known as GStore, where Google swag is sold) customer dataset to predict revenue per customer. Hopefully, the outcome will be more actionable operational changes and a better use of marketing budgets for those companies who choose to use data analysis on top of GA data.



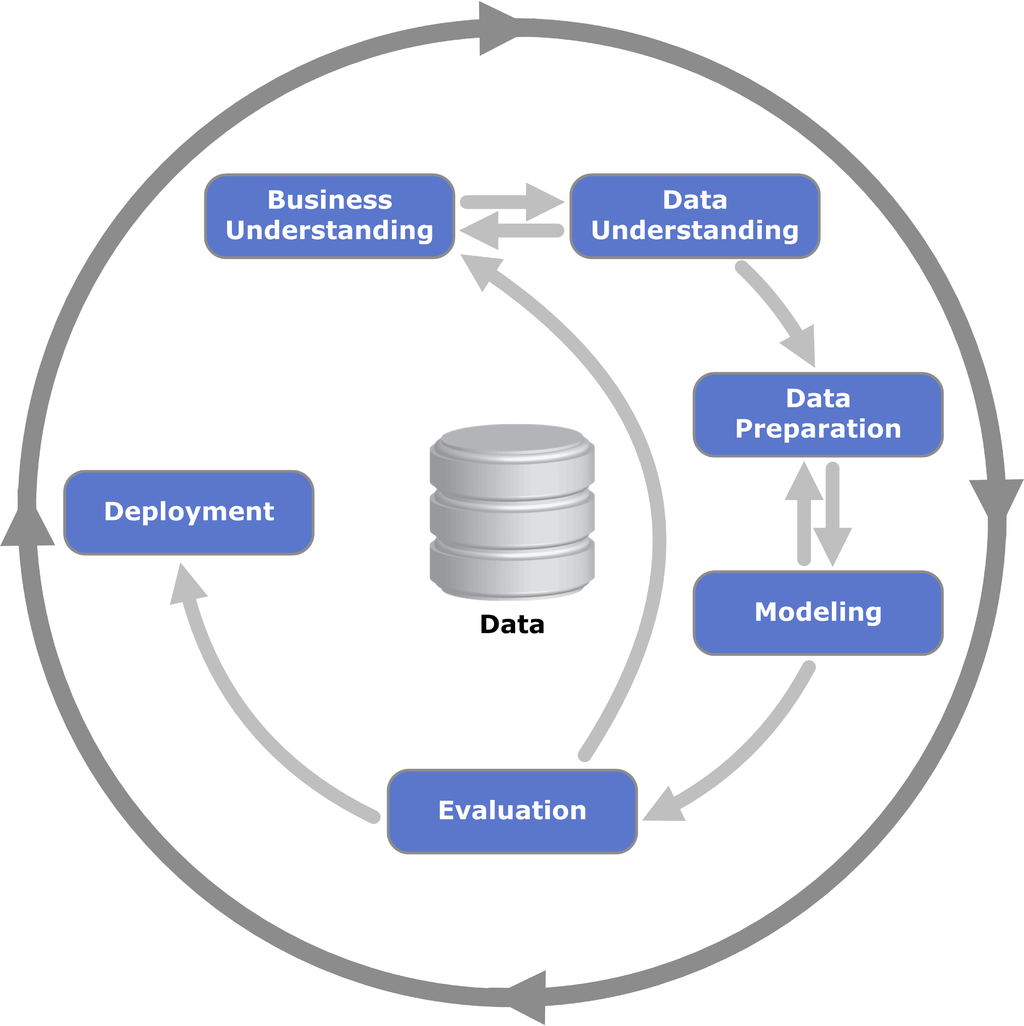
<https://www.googlemerchandisestore.com/>

Google Merchandise Store – Google Analytics

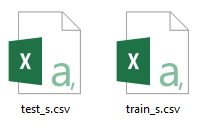


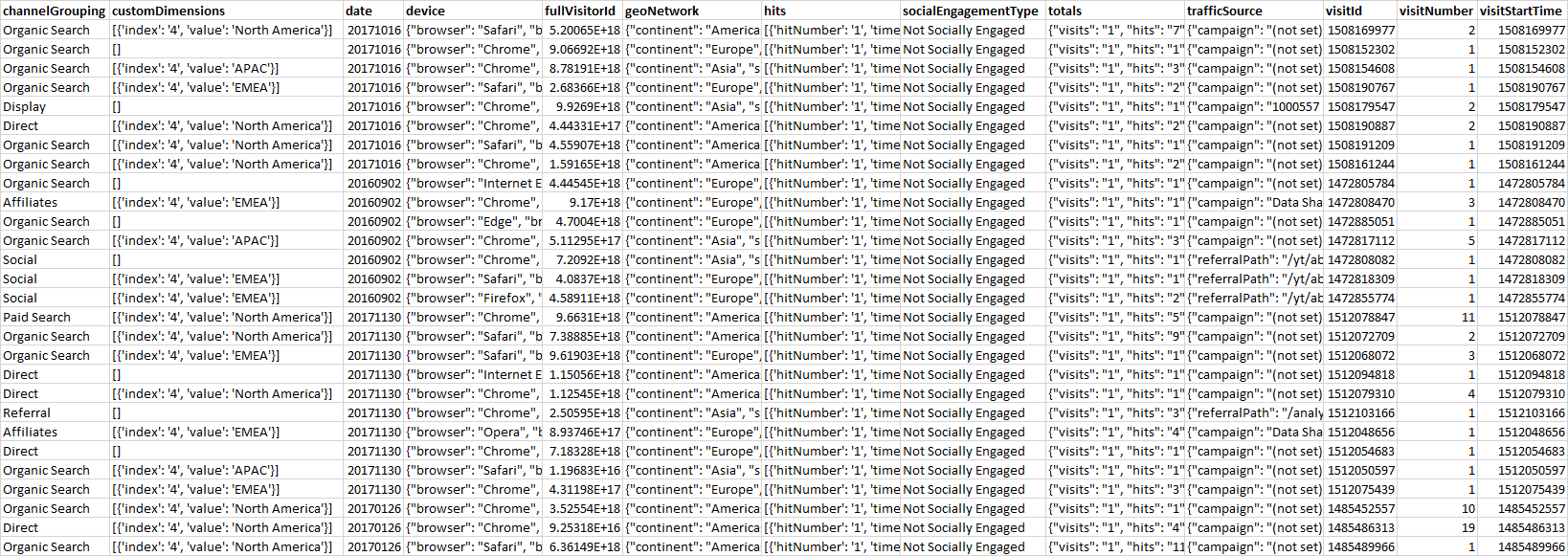
<https://analytics.google.com/analytics/web/demoAccount>

## Data analytics life cycle



### Data Understanding / Data Fields (Csv Files)

****

****

* ***channelGrouping*** - The channel via which the user came to the Store.
* ***customDimensions*** - This section contains any user-level or session-level custom dimensions that are set for a session. This is a repeated field and has an entry for each dimension that is set.
* ***date*** - The date on which the user visited the Store.
* ***device*** - The specifications for the device used to access the Store.
* ***fullVisitorId***- A unique identifier for each user of the Google Merchandise Store.
* ***geoNetwork*** - This section contains information about the geography of the user.
* ***hits*** - This row and nested fields are populated for any and all types of user behaviours (hits). Provides a record of all page visits.
* ***socialEngagementType*** - Engagement type, either "Socially Engaged" or "Not Socially Engaged".
* ***totals*** - This set of columns mostly includes high-level aggregate data, including target variable: transactionRevenue, e.g. one sample field value of ‘totals’:

**{**

**"visits" : "1",**

**"hits" : "16",**

**"pageviews" : "15",**

**"timeOnSite" : "225",**

**"transactions" : "1",**

**"transactionRevenue" : "21990000",**

**"newVisits" : "1",**

**"totalTransactionRevenue" : "29990000",**

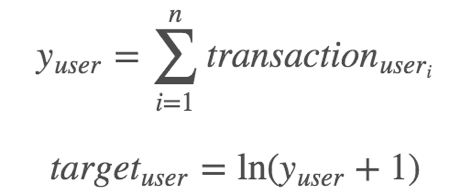
**"sessionQualityDim" : "6"**

**}**

* ***trafficSource*** - This section contains information about the Traffic Source from which the session originated.
* ***visitId*** - An identifier for this session. This is part of the value usually stored as the \_utmb cookie. This is only unique to the user. For a completely unique ID, you should use a combination of fullVisitorId and visitId.
* ***visitNumber*** - The session number for this user. If this is the first session, then this is set to 1.
* ***visitStartTime*** - The timestamp (expressed as POSIX time).

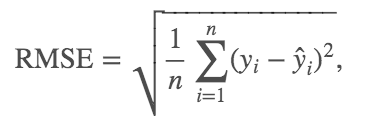
### What to predict ?

We are predicting the **natural log** of the sum of all transactions **per user**. For every user in the test set, the target is:



### Evaluation: Root Mean Squared Error (RMSE)

Submissions are scored on the root mean squared error. RMSE is defined as:

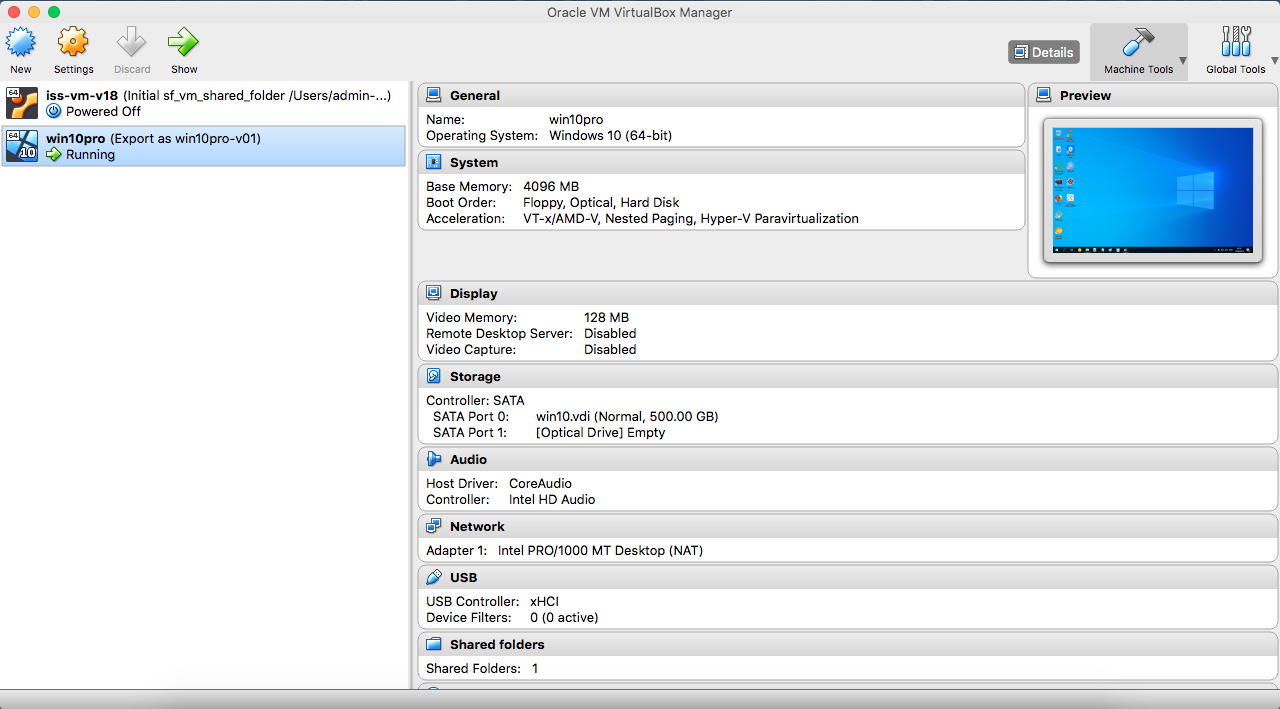


where y hat is the natural log of the predicted revenue for a customer and y is the natural log of the actual summed revenue value plus one.

## Analytics using software

### Download & install Software: ***VirtualBox 5.2.20***

<https://download.virtualbox.org/virtualbox/5.2.20/VirtualBox-5.2.20-125813-Win.exe>



### Download & install Software: ***VirtualBox 5.2.20 Extension Pack***

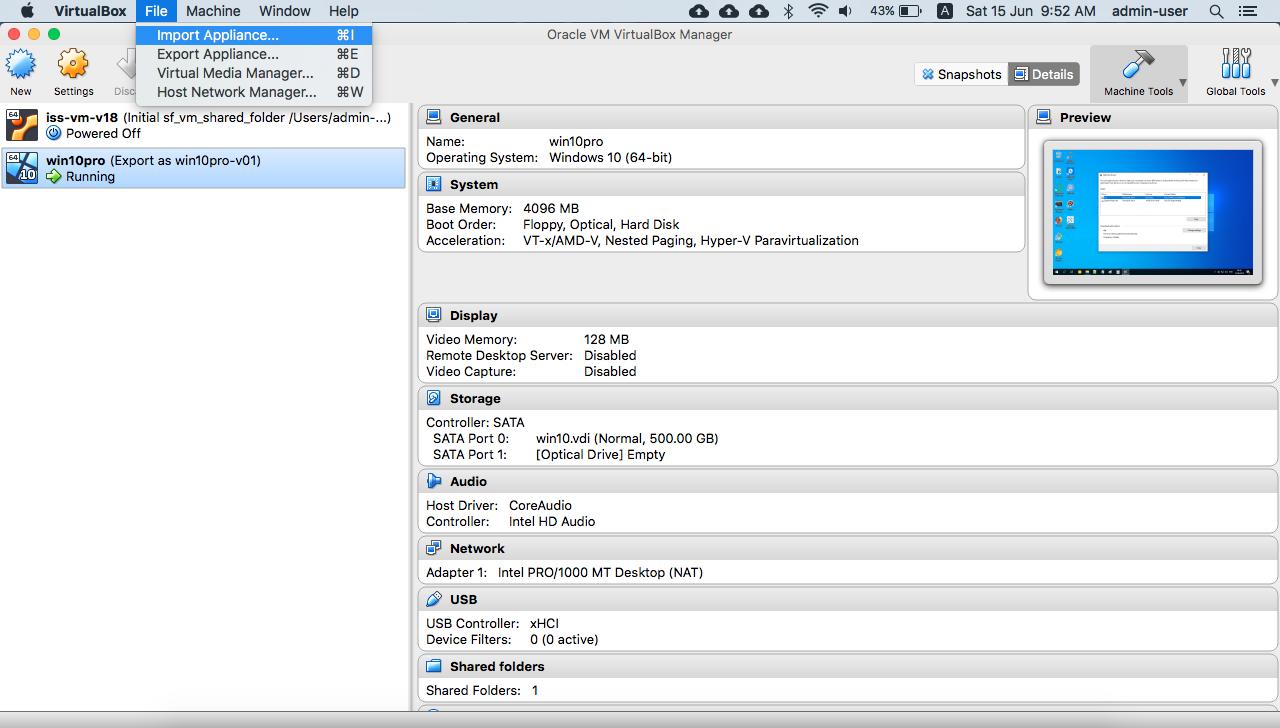
<https://download.virtualbox.org/virtualbox/5.2.20/Oracle_VM_VirtualBox_Extension_Pack-5.2.20.vbox-extpack>

### Download virtual machine workstation: ***win10pro***

$$$link

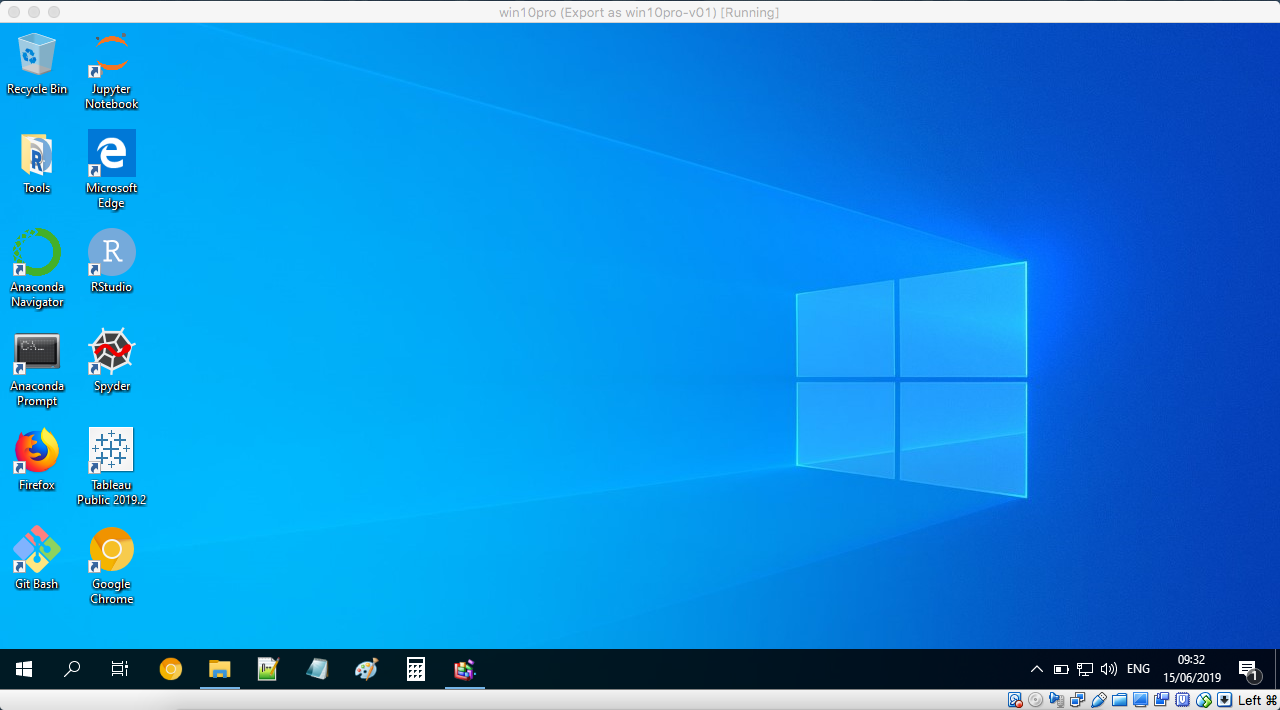
### Import & run virtual machine workstation: ***win10pro***

Start VirtualBox, click ***File 🡪 Import Appliance***.



Select the virtual machine image/file: ***win10pro.ova***

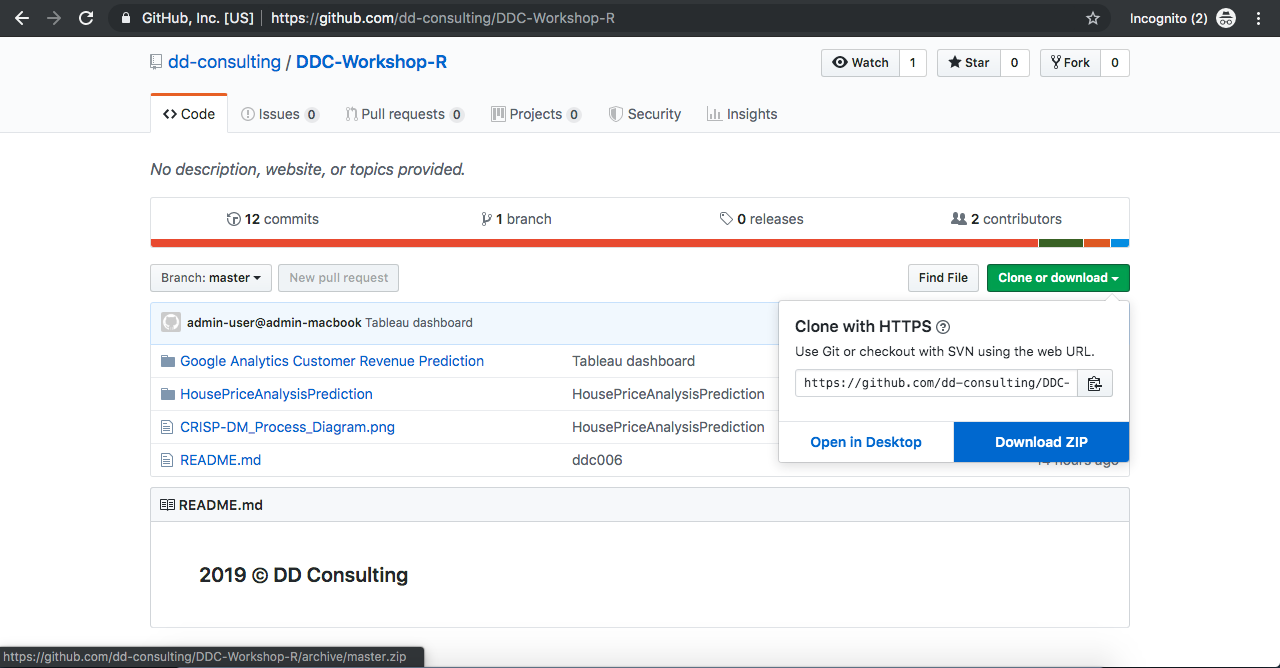
Click show  to start virtual machine.



### Download workshop materials

In win10pro virtual machine, open web browser: <https://github.com/dd-consulting/DDC-Workshop-R>

Click **Download Zip** button.



### Extract/Unzip workshop materials to ***download*** folder

$

### from desktop, Strat software: ***rstudio***

$

### set working directory

$

### Open R scrpt file

E.g. Google Analytics Customer Revenue Prediction/code/r-eda-for-gstore-glm-xgb.Rmd

### Run R scripts: EDA & Predictive Modelling

$

$

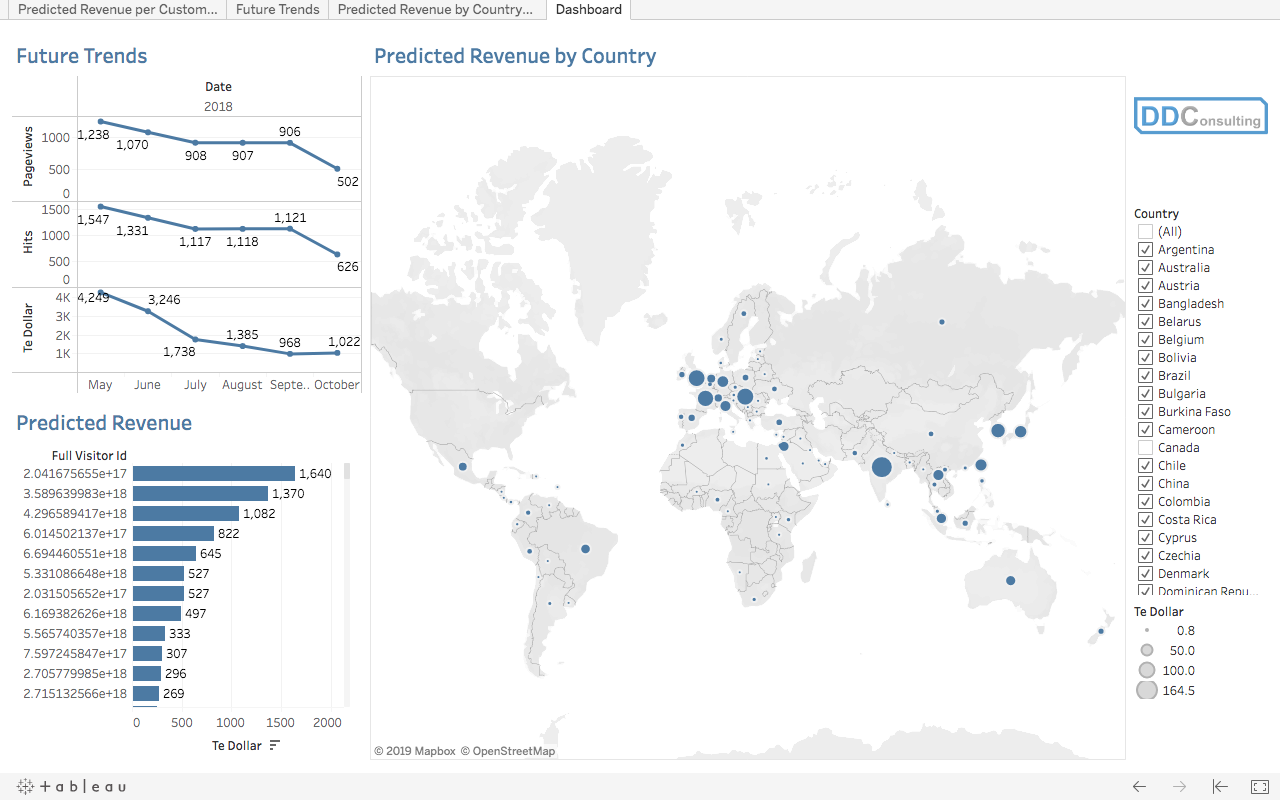
### from desktop, Strat software: ***tableau***

$

### Open Tableau file: Interactive Data visualization

E.g. Google Analytics Customer Revenue Prediction/Tableau dashboard/DDC - Google Analytics Customer Revenue Prediction.twbx

### explore, edit, or Create Tableau visualizations



## Summary

What we have experienced/learnt:

* Analyse a case of customer revenue for Google Merchandise Store, using Google Analytics data.

What’s the business value?

* Understand the various features available from Google Analytics data.

What are some example features?

* Able to conduct exploratory data analysis (EDA) using ***R***

What are some example EDA charts?

* Able to use predictive models to forecast future customer revenue using ***R***

What’s the predictive models/algorithms we used?

* Able to identify important features which can indicate customer purchases (data insights)

What are some example data insights?

* Able to build interactive data visualization dashboard to present results & insights using ***Tableau***

What are dimension and measure?



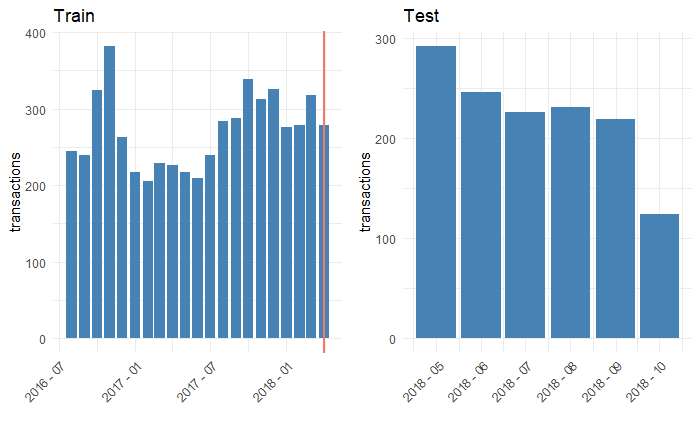
## Workshop Exercise / Quiz

Question 1 : How many **months** of data are there in training data?

Your answer :

Question 2 : How many **months** of data are there in testing data?

Your answer :



Question 3 : What’s the **variable name** of the target customer revenue we want to predict?

Your answer :

﻿ [1] ﻿"channelGrouping" "date"

[3] "fullVisitorId" "visitId"

[5] "visitNumber" "visitStartTime"

[7] "browser" "operatingSystem"

[9] "isMobile" "deviceCategory"

[11] "continent" "subContinent"

[13] "country" "region"

[15] "metro" "city"

[17] "networkDomain" "campaign"

[19] "source" "medium"

[21] "keyword" "isTrueDirect"

[23] "adContent" "referralPath"

[25] "adwordsClickInfo.page" "adwordsClickInfo.slot"

[27] "adwordsClickInfo.gclId" "adwordsClickInfo.adNetworkType"

[29] "adwordsClickInfo.isVideoAd" "hits1"

[31] "pageviews" "timeOnSite"

[33] "sessionQualityDim" "newVisits"

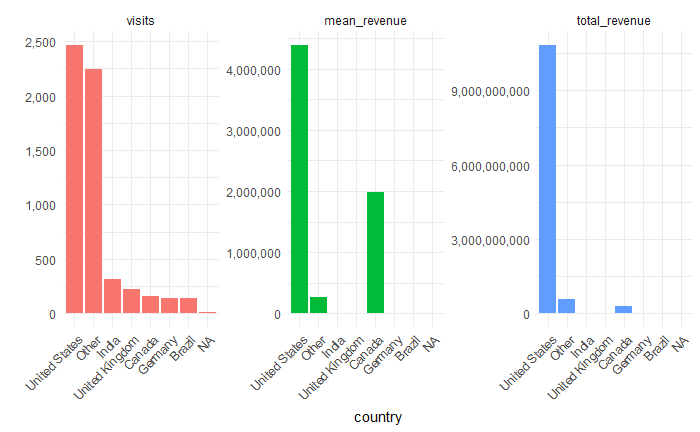
[35] "bounces" "transactionRevenue"

Question 4 : What are the **top three** countries interested in Google swag/products?

Your answer :

Question 5 : On average, which country’s customer spends **most**?

Your answer :

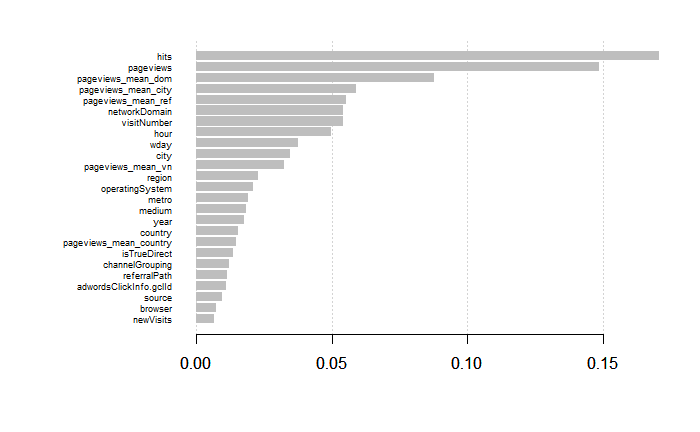


Question 6 : What are the **three** **most** influential indicators for revenue prediction?

Your answer :

Question 7 : What are the **three** **least** influential indicators for revenue prediction?

Your answer :



$ channelGrouping <fct> Organic Search, Organic Search, Organic Search, Organic ...

$ visitNumber <int> 2, 1, 1, 1, 1, 2, 2, 1, 1, 1, 3, 1, 5, 2, 1, 1, 1, 1, 11...

$ browser <fct> Safari, Chrome, Chrome, Chrome, Safari, Chrome, Chrome, ...

$ operatingSystem <fct> iOS, Windows, Windows, Windows, Macintosh, Android, Wind...

$ isMobile <int> 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

$ deviceCategory <fct> mobile, desktop, desktop, desktop, desktop, mobile, desk...

$ continent <fct> Americas, Americas, Europe, Asia, Europe, Asia, Americas...

$ subContinent <fct> Northern America, Northern America, Southern Europe, Sou...

$ country <fct> Canada, Canada, Portugal, India, United Kingdom, Saudi A...

$ region <fct> NA, NA, NA, NA, NA, Riyadh Province, New York, NA, New Y...

$ metro <fct> NA, NA, NA, NA, NA, NA, New York NY, NA, New York NY, NA...

$ city <fct> NA, NA, NA, NA, NA, Riyadh, New York, NA, New York, NA, ...

$ networkDomain <fct> NA, NA, vodafone.pt, NA, as9105.com, NA, NA, verizon.net...

$ campaign <fct> NA, NA, NA, NA, NA, 1000557 | GA | US | en | Hybrid | GD...

$ source <fct> google, google, google, google, google, google, (direct)...

$ medium <fct> organic, organic, organic, organic, organic, cpc, NA, or...

$ keyword <fct> NA, NA, NA, NA, NA, (User vertical targeting), NA, NA, N...

$ isTrueDirect <int> 1, NA, NA, NA, NA, NA, 1, NA, NA, NA, NA, NA, 1, NA, NA,...

$ adContent <fct> NA, NA, NA, NA, NA, Google Merchandise Store, NA, NA, NA...

$ referralPath <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...

$ adwordsClickInfo.page <fct> NA, NA, NA, NA, NA, 1, NA, NA, NA, NA, NA, NA, NA, 1, NA...

$ adwordsClickInfo.slot <fct> NA, NA, NA, NA, NA, RHS, NA, NA, NA, NA, NA, NA, NA, Top...

$ adwordsClickInfo.gclId <fct> NA, NA, NA, NA, NA, CL2-\_8Pm9dYCFU9MDQodfdgCIg, NA, NA, ...

$ adwordsClickInfo.adNetworkType <fct> NA, NA, NA, NA, NA, Content, NA, NA, NA, NA, NA, NA, NA,...

$ adwordsClickInfo.isVideoAd <int> NA, NA, NA, NA, NA, 0, NA, NA, NA, NA, NA, NA, NA, 0, NA...

$ pageviews <int> 7, 14, 1, 3, 2, 1, 2, 1, 2, 1, 1, 1, 3, 7, 16, 1, 1, 2, ...

$ newVisits <int> NA, 1, 1, 1, 1, NA, NA, 1, 1, 1, NA, 1, NA, NA, 1, 1, 1,...

$ bounces <int> NA, NA, 1, NA, NA, 1, NA, 1, NA, 1, 1, 1, NA, NA, NA, 1,...

$ hits <int> 7, 18, 1, 3, 2, 1, 2, 1, 2, 1, 1, 1, 3, 7, 21, 1, 1, 2, ...

$ year <fct> 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 20...

$ wday <fct> 2, 2, 2, 2, 2, 2, 2, 2, 2, 6, 6, 6, 6, 6, 6, 6, 6, 6, 5,...

$ hour <fct> 16, 2, 11, 11, 21, 18, 21, 22, 13, 8, 9, 6, 11, 4, 0, 9,...

$ pageviews\_mean\_vn <dbl> 4.493438, 3.476790, 3.476790, 3.476790, 3.476790, 4.4934...

$ pageviews\_mean\_country <dbl> 5.619792, 5.619792, 3.370370, 2.926108, 2.663194, 3.1764...

$ pageviews\_mean\_city <dbl> 7.000000, 14.000000, 1.000000, 3.000000, 2.000000, 1.000...

$ pageviews\_mean\_dom <dbl> 7.000000, 14.000000, 5.500000, 3.000000, 1.571429, 1.000...

$ pageviews\_mean\_ref <dbl> 7.000000, 14.000000, 1.000000, 3.000000, 2.000000, 1.000...

## Your Valuable Feedback

**Your Full Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**How likely would you be to recommend this learning opportunity to a friend, classmate or colleague?**

○ Unlikely ○ Maybe ○ Very likely ○ Certainly

**How much proportion of this learning/workshop could you follow and understand?**

○ < 25% ○ 25 ~ 50% ○ 50 ~ 75% ○ > 75%

**How much new information/knowledge/skill did you experience in this learning session?**

○ Nothing new ○ Some ○ Majority ○ Almost all

**How engaging was the presenter: Sam?**

○ Unimpressive ○ Sort of ○ Good ○ Highly engaging

**Please tell us what topics you would like to learn more?**